TOSHIBA
Leading Innovation >>>

PPC64 RTパッチ性能計測

― アップデート版 ―

東芝 ソフトウェア技術センター
2007年 2月 22日

Copyright 2007, Toshiba Corporation.

目次 – 前回と同じ内容はタイトルを灰色にしています

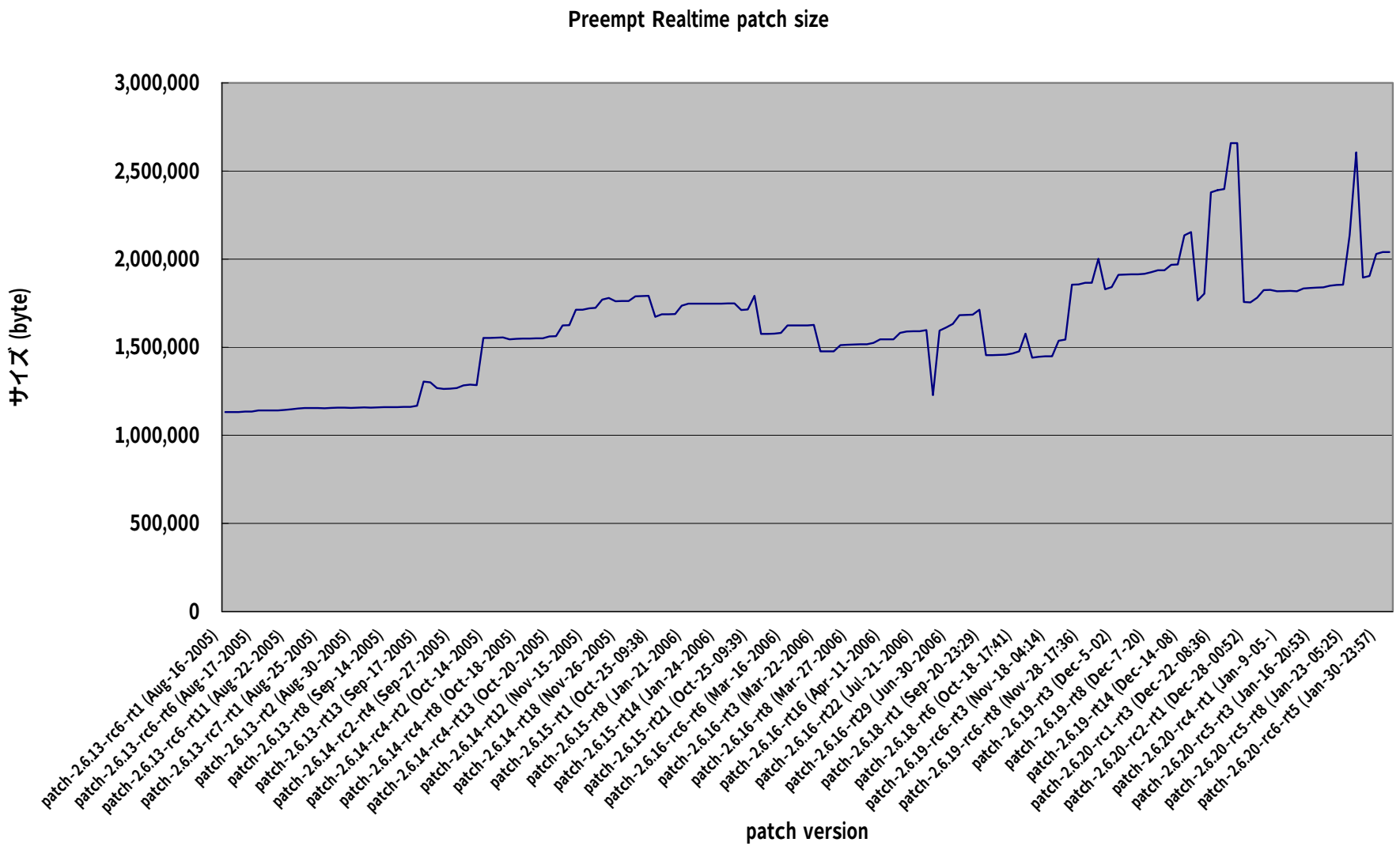
- **目的**
- **Cellリファレンスセット**
- **計測方法**
- **計測結果**
- **まとめ**

realtime preempt patch

- **Ingo Molnar氏らによるリアルタイム性向上のためのパッチ**
  - Linux Kernel スケジューラのmaintainer
  - Thomas Gleixner氏、他
  - 2004年頃から。mainlineに取り込また部分もある
  - Kernel内にpreemption point を追加
    - spinlock → mutex w/ priority inheritance
    - hard/soft interrupts → kernel threads
    - 割り込み禁止区間の短縮
    - etc.
  - x86_64, i386, arm, mips, powerpc(32), …
  - http://people.redhat.com/mingo/realtime-preempt/

- **関連URL**
  - http://rt.wiki.kernel.org/
  - http://tree.celinuxforum.org/CelfPubWiki/JapanTechnicalJamboree11
  - http://opentechpress.jp/developer/article.pl?sid=05/01/30/227247
  - http://www.atmarkit.co.jp/fembedded/rtos02/rtos02c.html

realtime preempt パッチサイズ



目的

- **PPC64 RTパッチの有効性検討**
  - **実装はしてみたが、効果の程は？　計ってみよう！**
  - **対象**：NonRT**版** kernel vs. RT**版** kernel
  - **目安**：**応答性能** vs. **処理性能**
  - **応答性能**：lpar**割込み遅延時間**

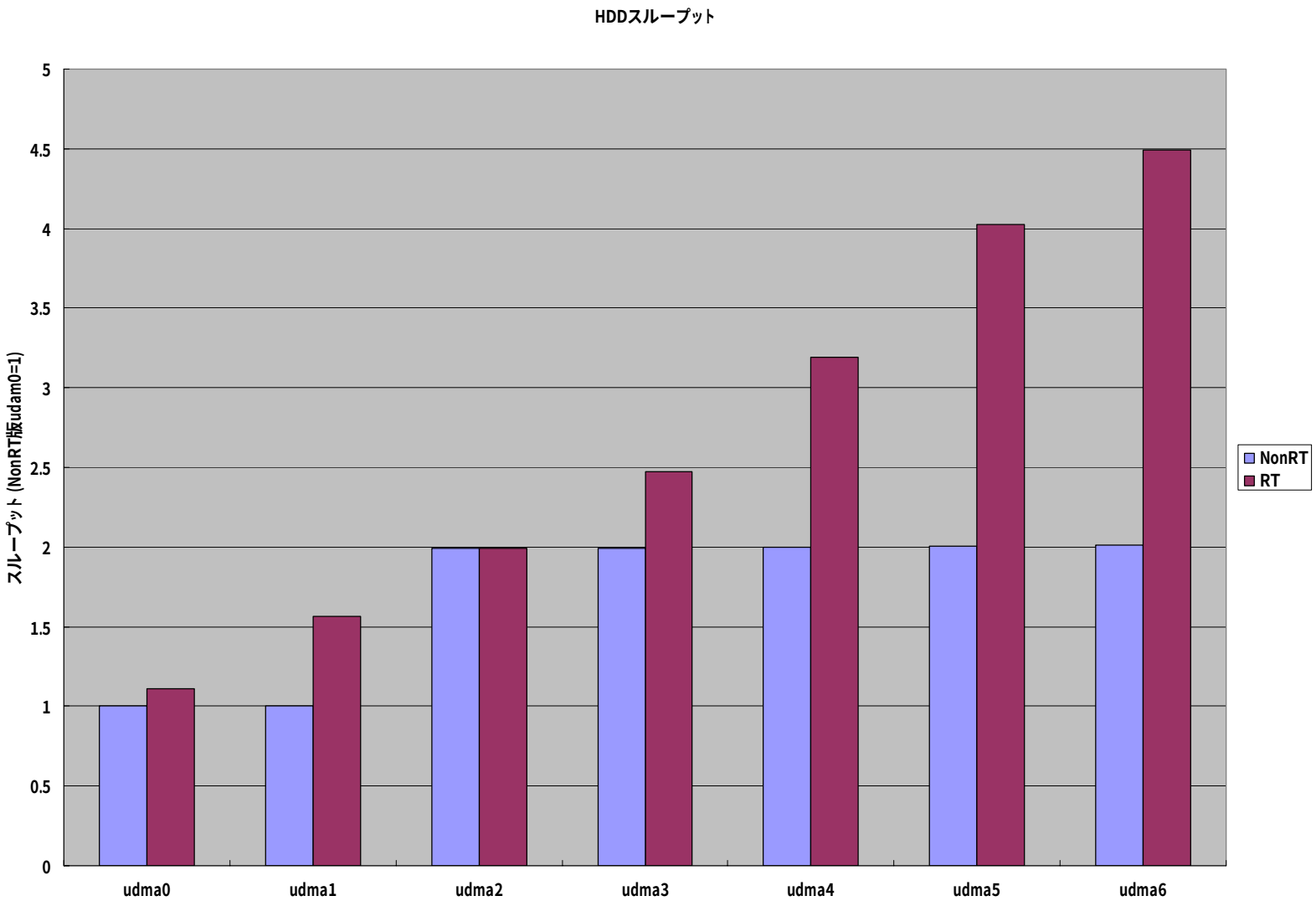
    ping round trip time
  - **処理性能**：netperf, hdparm

今回の目的

- **申し訳ありません！**
  - **カーネルにバグを入れ込んでいました！**
  - PowerPC**へのRTパッチ移植に関するミスです**
    - **影響はNonRT版のみ (と思われます)**

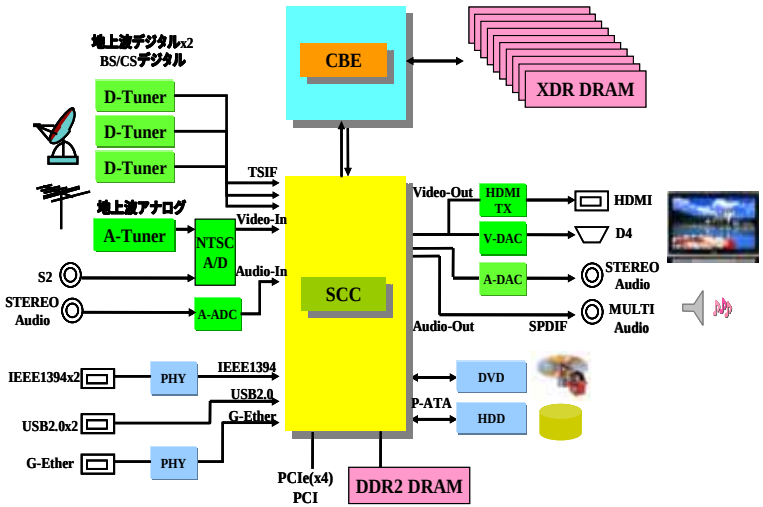
- **今回、再計測した結果をご報告します。**

HDD性能例 (不具合のあったカーネル)



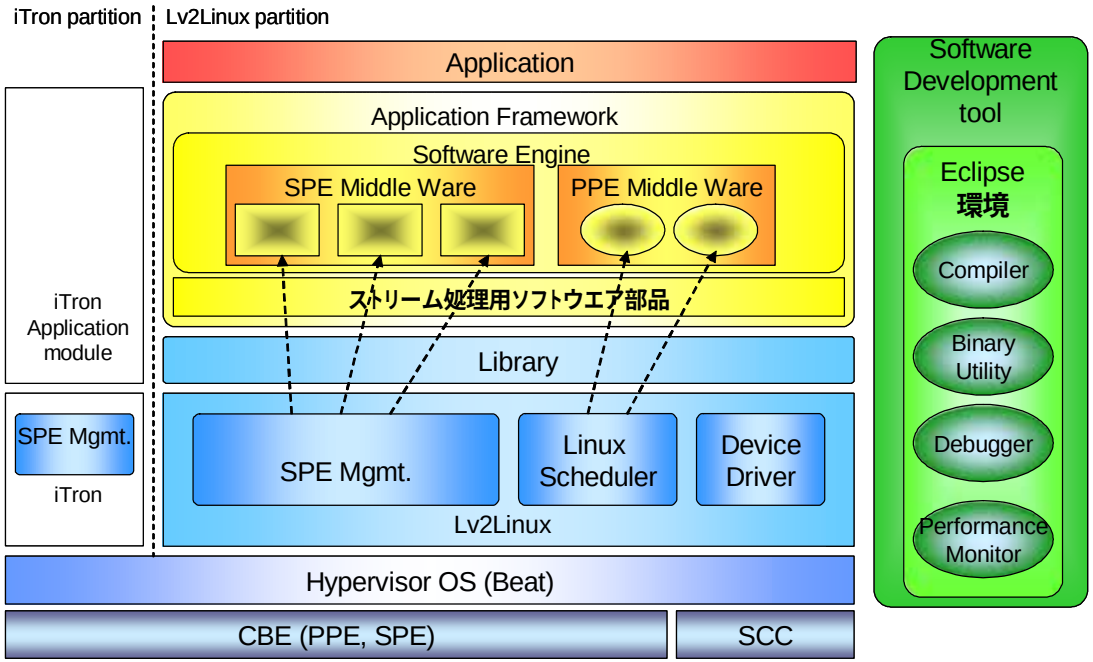
Cellリファレンスセット

- http://www.toshiba.co.jp/tech/review/2006/06
- http://www.semicon.toshiba.co.jp/product/micro/cell/reference.html
- **HW**



- **SW**
  - **ハードウェア資源を管理するハイパーバイザ**OS
  - Lv2Linux (**論理パーティション**lpar**上の**Linux)
    - 2.6.12.3 kernel
    - 2.6.12.final-V0.7.51-33.patch (http://people.redhat.com/mingo/realtime-preempt)
      - **既に削除されていますが**…

CellリファレンスセットSW構成



◆基本ソフトウェアとして Hypervisor OS「Beat」、ゲストOSとして「Lv2Linux」を提供

◆SPEの効率的利用を可能にするSPEマネージメント

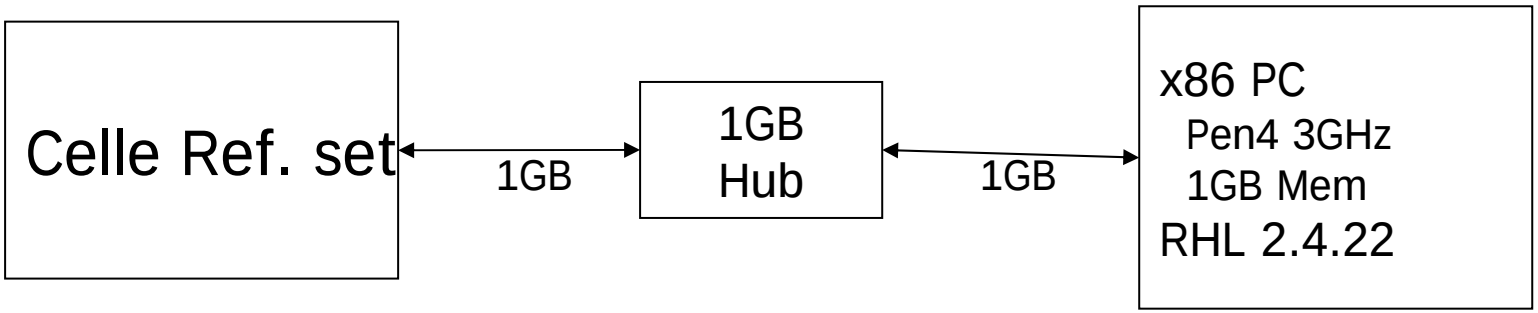
◆SCCの多様なIOに対応したデバイスドライバ

◆Eclipse統合開発環境：コンパイラ、PPE/SPE統合デバッガ、パフォーマンスモニタ,etc.

カーネルコンフィギュレーション

- **次の2種類のカーネルで測定**
  - NonRT**版**
    - PREEMPT_NONE=y
    - #define PPC64_TLB_BATCH_NR 1
  - RT**版**
    - PREEMPT_RT=y
    - PREEMPT**系のデバッグオプションなし**
    - #define PPC64_TLB_BATCH_NR 1
  - **一部、**NonRT**版** + PREEMPT_{HARD,SOFT}IRQS=y

計測環境



RT性能計測方法

- **lpar割込み応答性能**
  - **論理パーティション切替え時に発生する割り込みを利用**
  - **周期的にlpar割込み発生**
  - **ハンドラ到着時刻とlpar切替え基点との差分**
- **ping応答性能**
  - *“ping –f –c 100 [-s 4000 | -s 8192] <addr>”*
    - **出力の** rtt min/avg/max **を計測**
    - **受信のみ**
      - Busybox v1.01 ping **には** “-f” option **がなかった**
- **処理性能**
  - **ネットワーク** *“netpef –H <addr>”, v*2.4.1
  - **HDD** *“hdparm –t /dev/had”, busybox v1.01*

RT性能計測

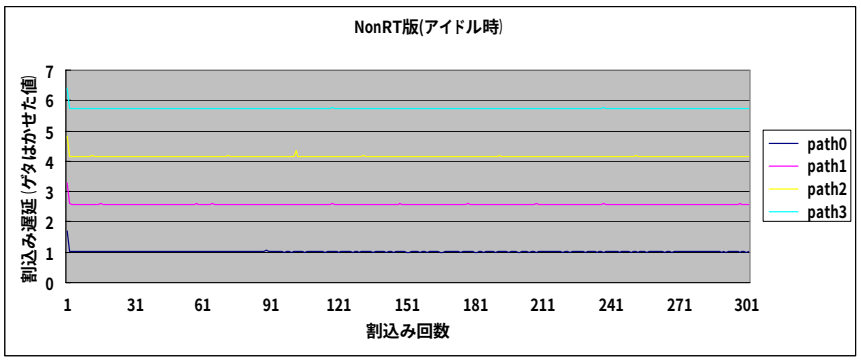
- **lpar 割込み応答性能**

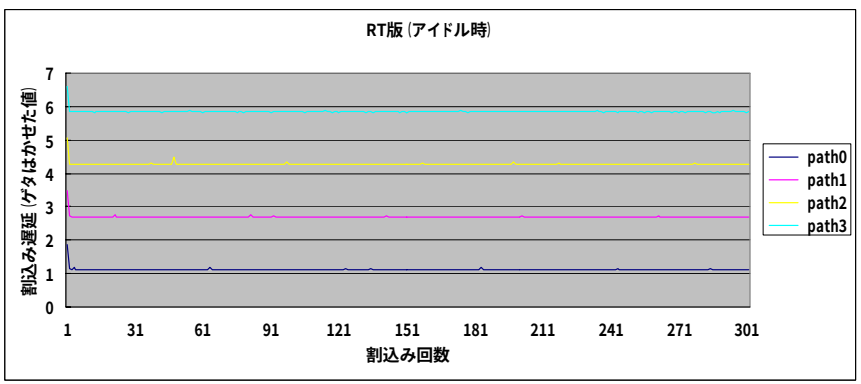
- **ping 応答性能**

- **処理性能**
  - **ネットワーク** netperf
  - HDD hdparm

Ipar割込み応答性能 ー NonRT&RT版（アイドル時）

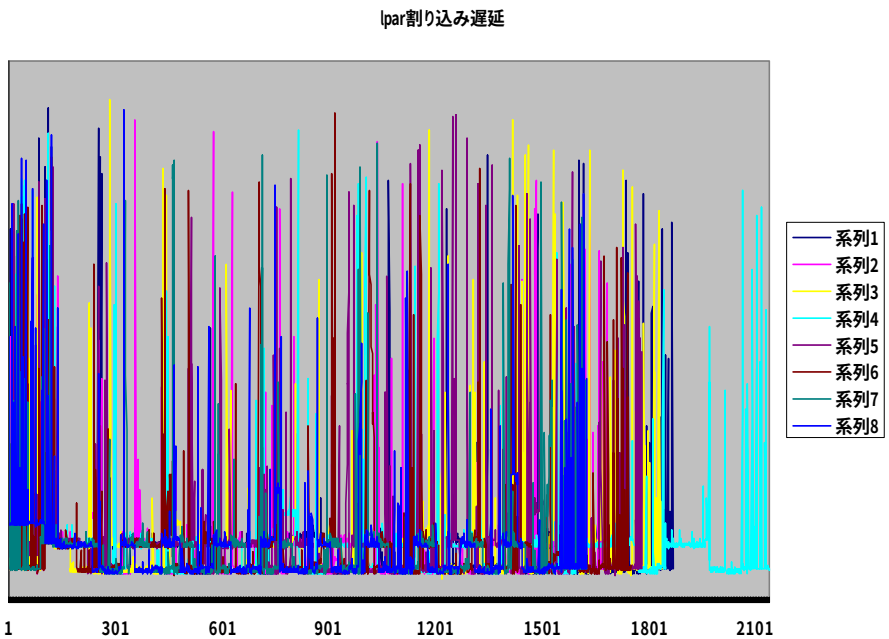
- **負荷のない状態では差異は見られず**



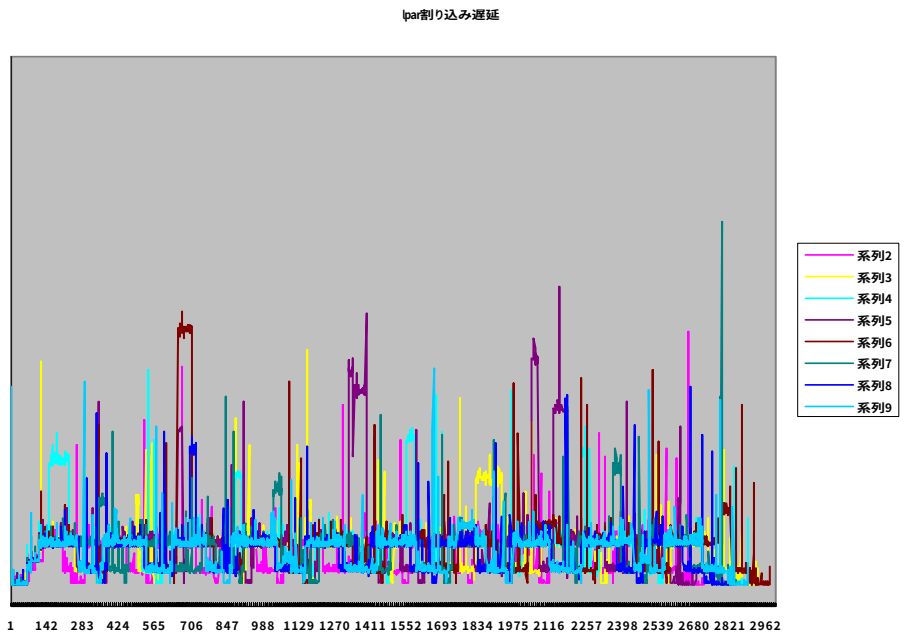


前回のIpar割込み応答性能（負荷あり）

NonRT版

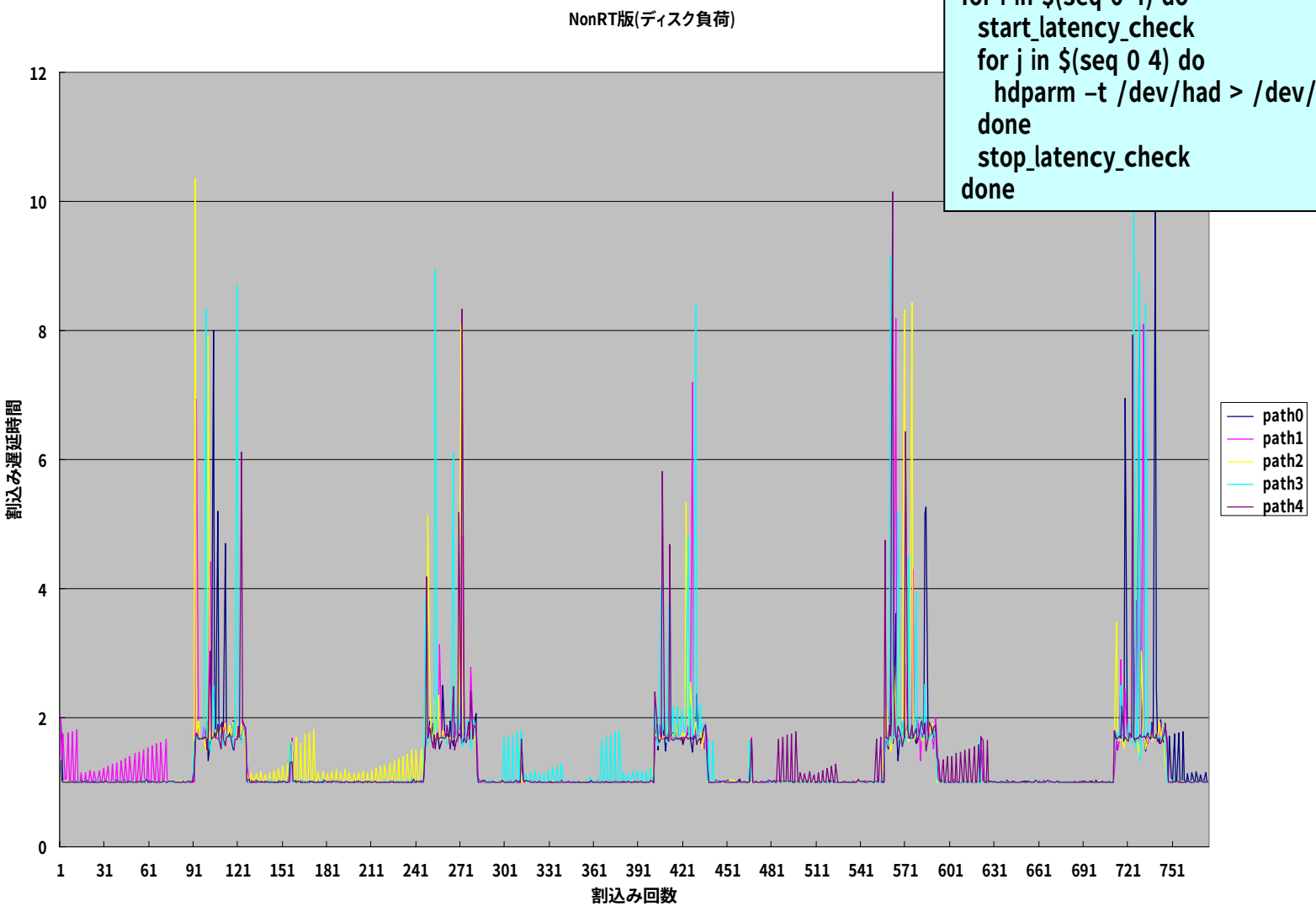


RT版

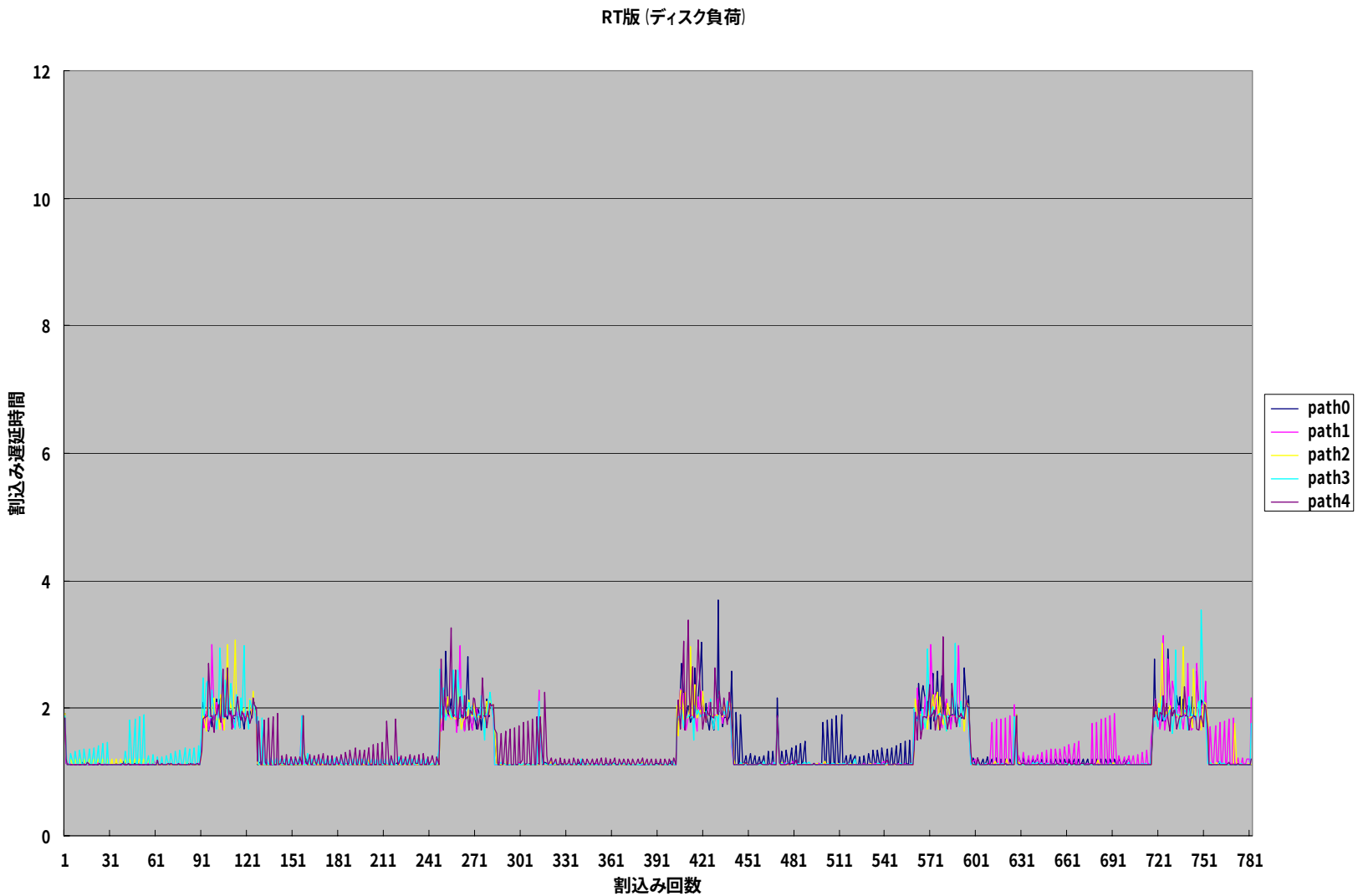


割込み遅延時間 ー NonRT版 (ディスク負荷時)

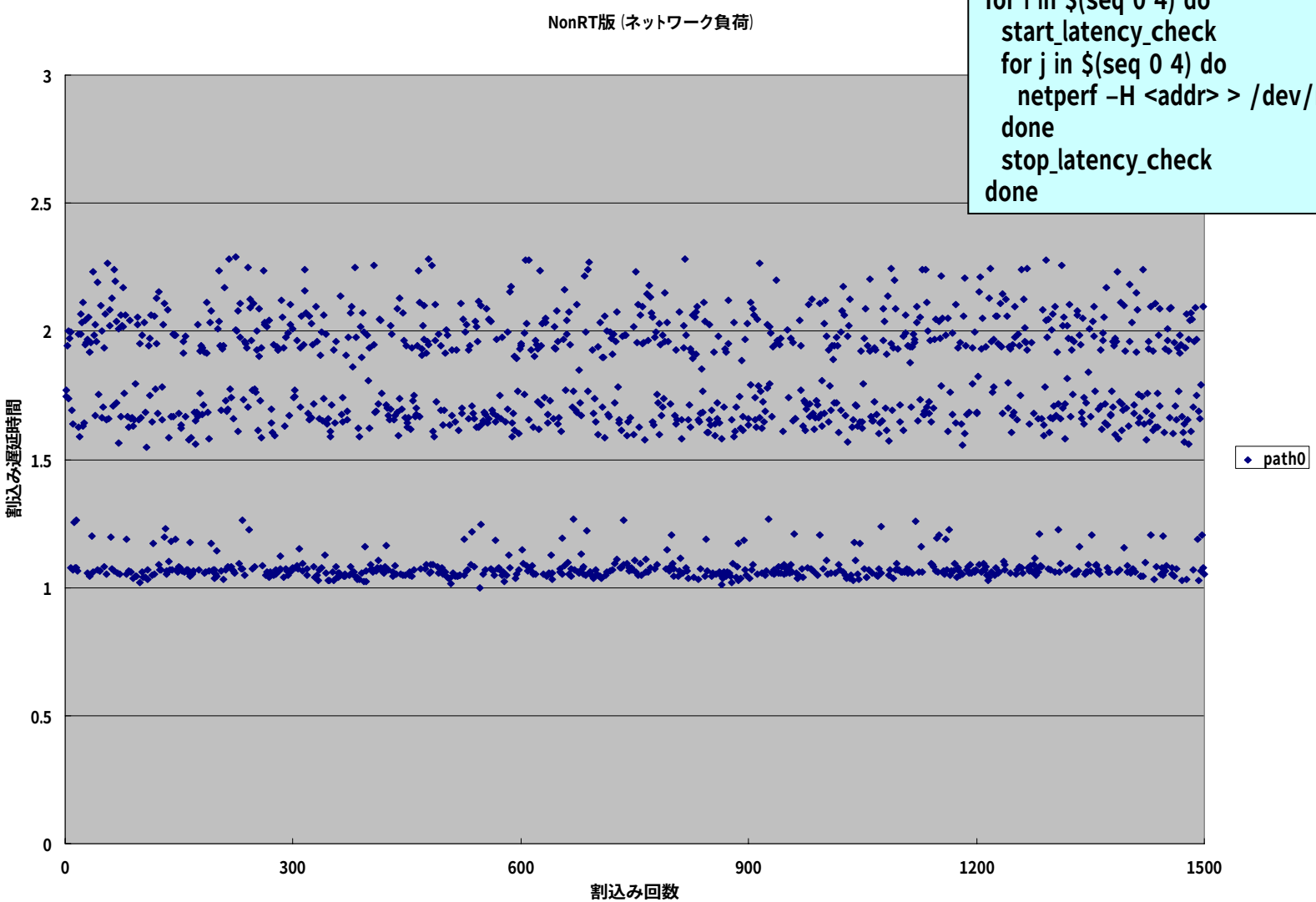
```
for i in $(seq 0 4) do
  start_latency_check
  for j in $(seq 0 4) do
    hdparm –t /dev/had > /dev/null
  done
  stop_latency_check
done
```



割込み遅延時間 ー RT版 (ディスク負荷時)

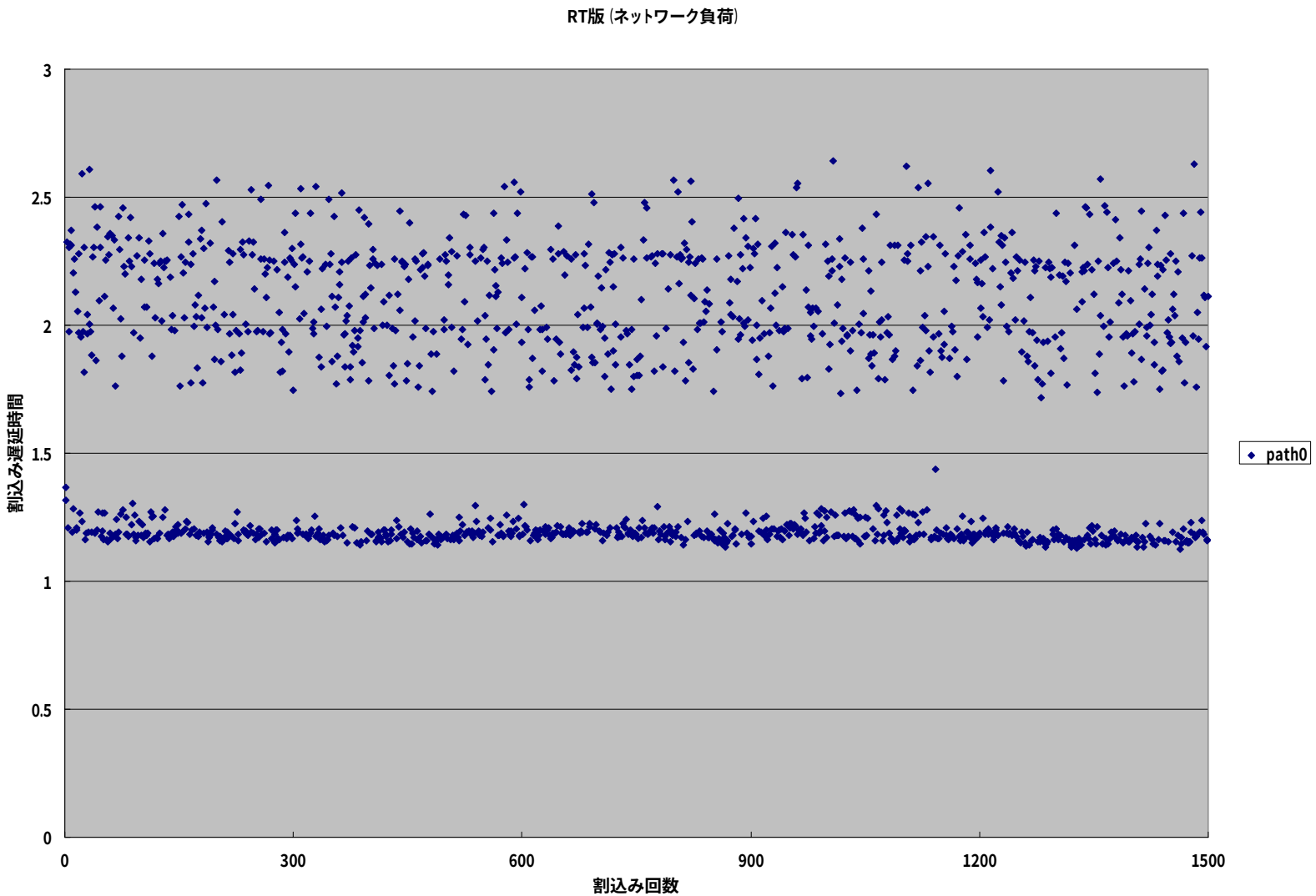


割込み遅延時間 ー NonRT版 (ネットワーク負荷時)

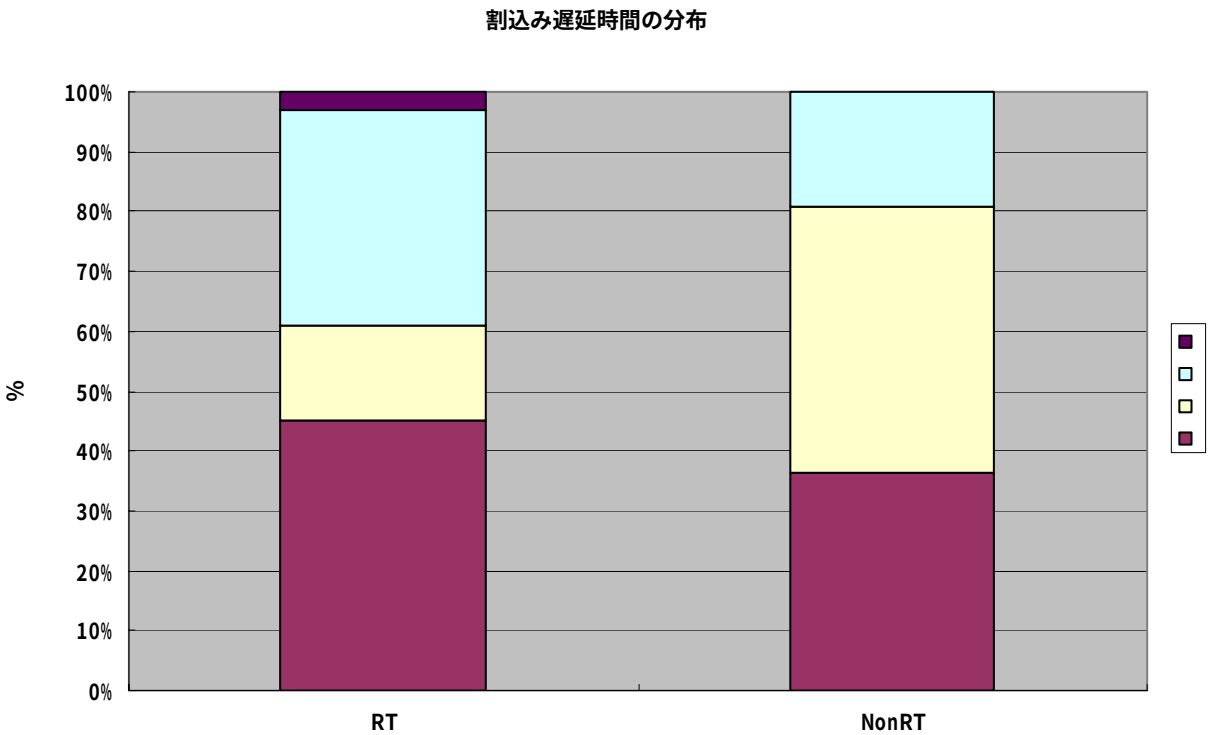



```
for i in $(seq 0 4) do
  start_latency_check
  for j in $(seq 0 4) do
    netperf –H <addr> > /dev/null
  done
  stop_latency_check
done
```

割込み遅延時間 ー RT版 (ネットワーク負荷時)



割込み遅延時間 ー ヒストグラム (ネットワーク負荷時)



RT性能計測

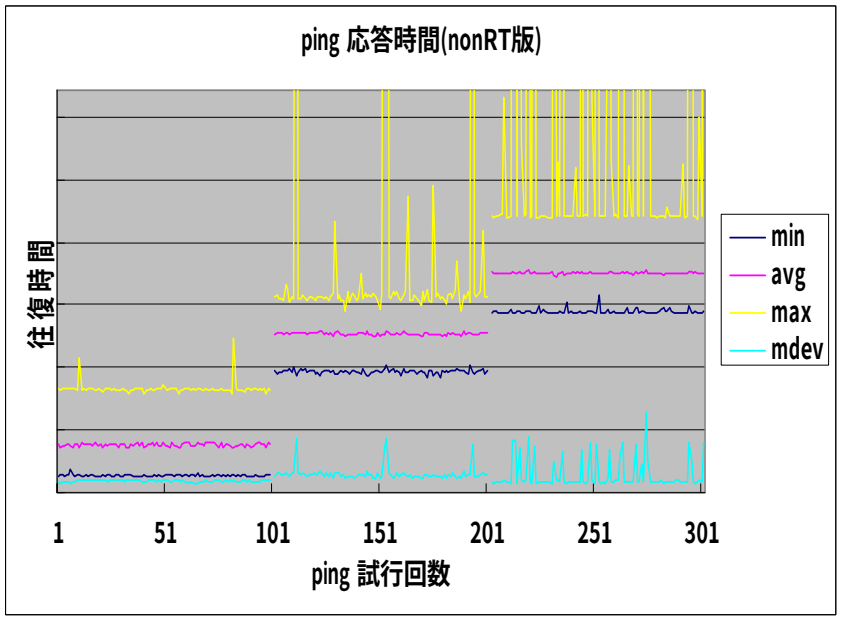
- **lpar 割込み応答性能**

- **ping 応答性能**

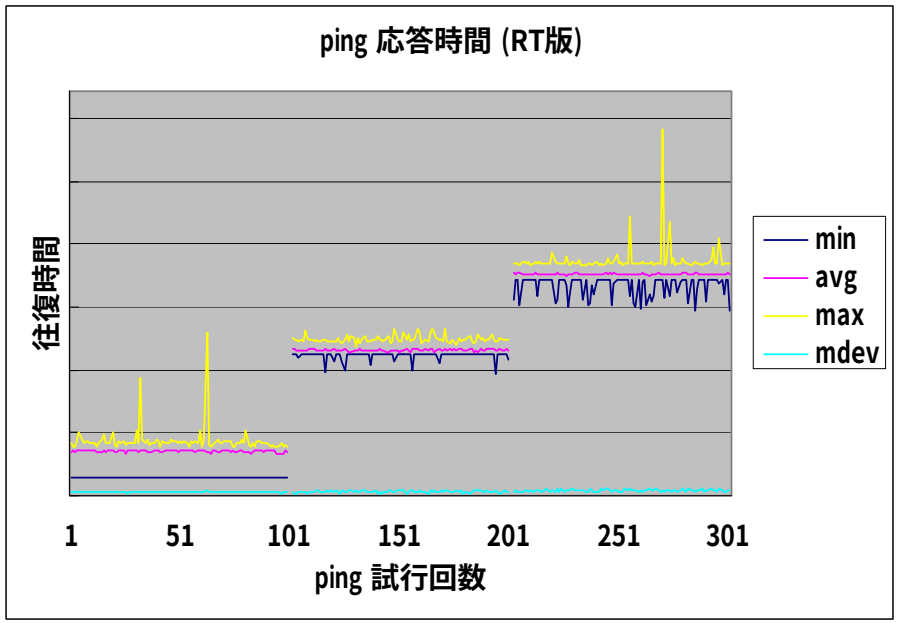
- **処理性能**
  - **ネットワーク** netperf
  - HDD hdparm

前回の ping –f 測定結果

NonRT版



RT版



- **試行回数**1-100：ping –f –c 1000
- 101-200：ping –f –c 1000 –s 4000
- 201-300：ping –f –c 1000 –s 8192

ping –f 測定結果

- **今回はお出しできるデータがありません。**

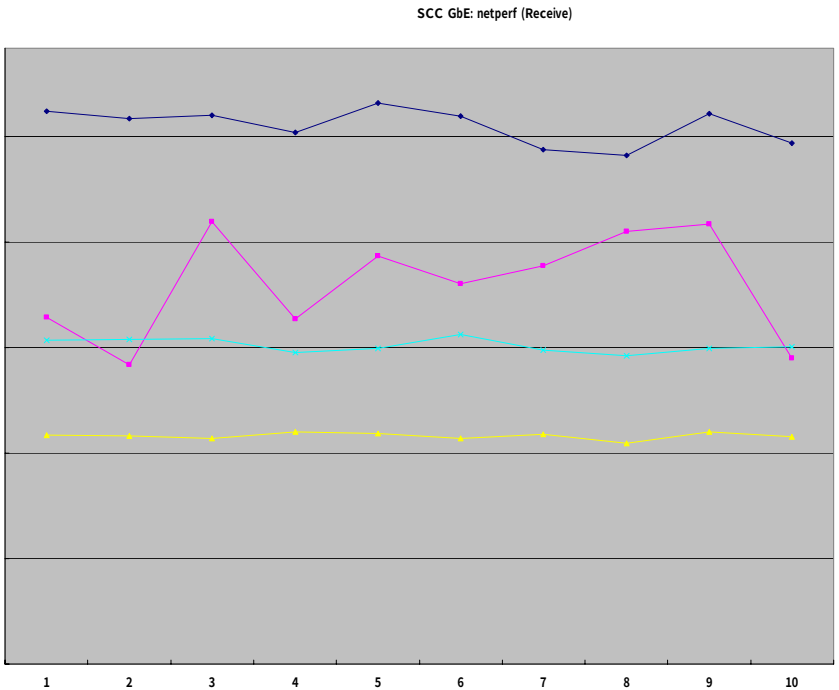
RT性能計測

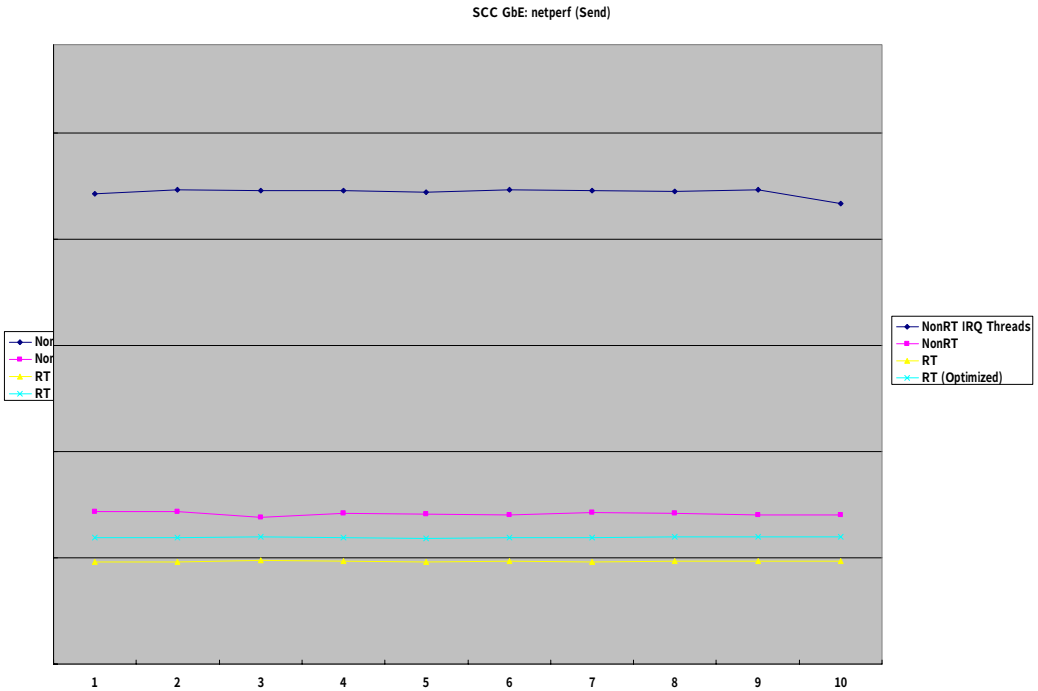
- **lpar 割込み応答性能**

- **ping 応答性能**

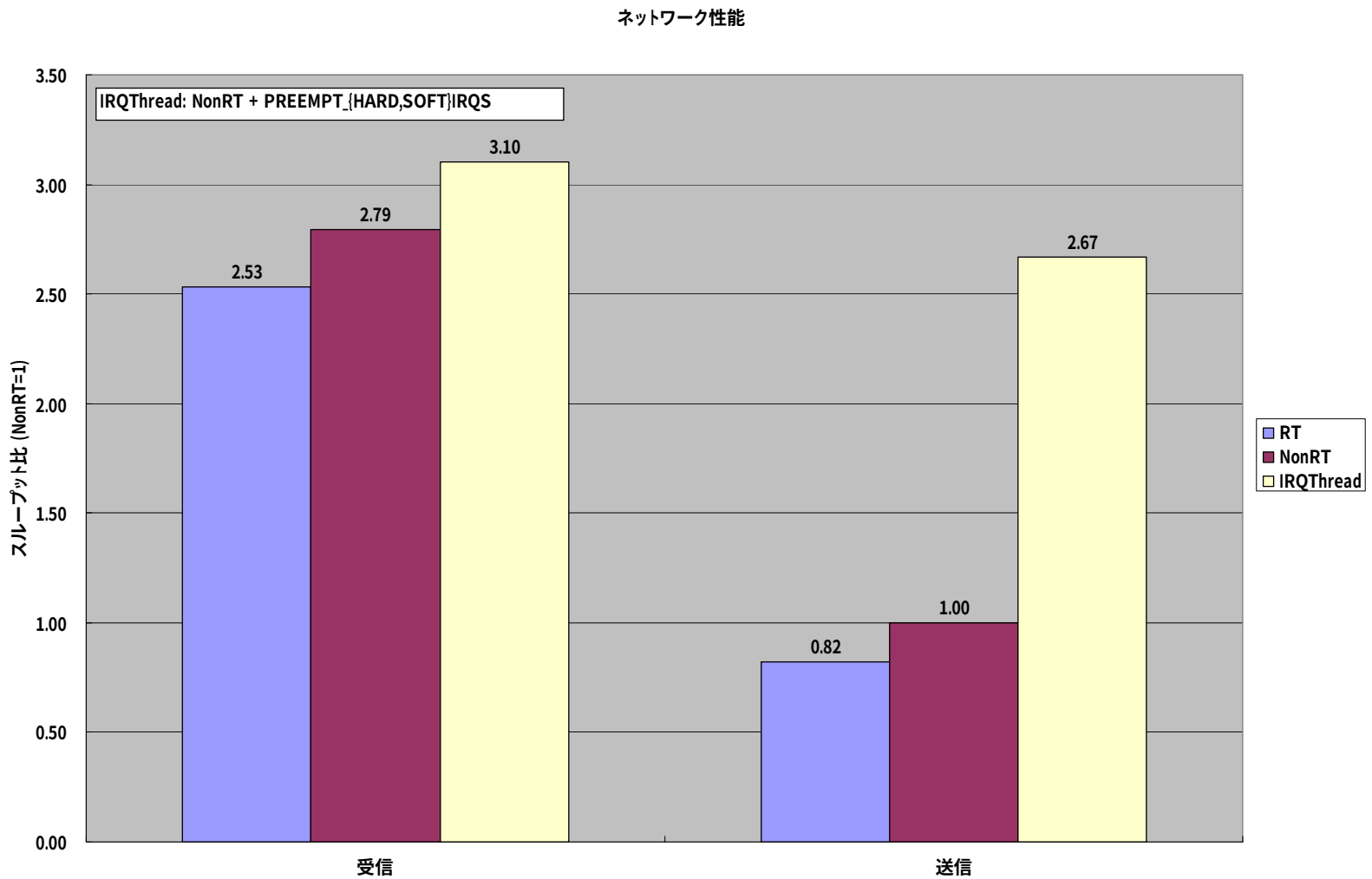
- **処理性能**
  - **ネットワーク** netperf
  - HDD hdparm

前回のスループット ー ネットワーク性能

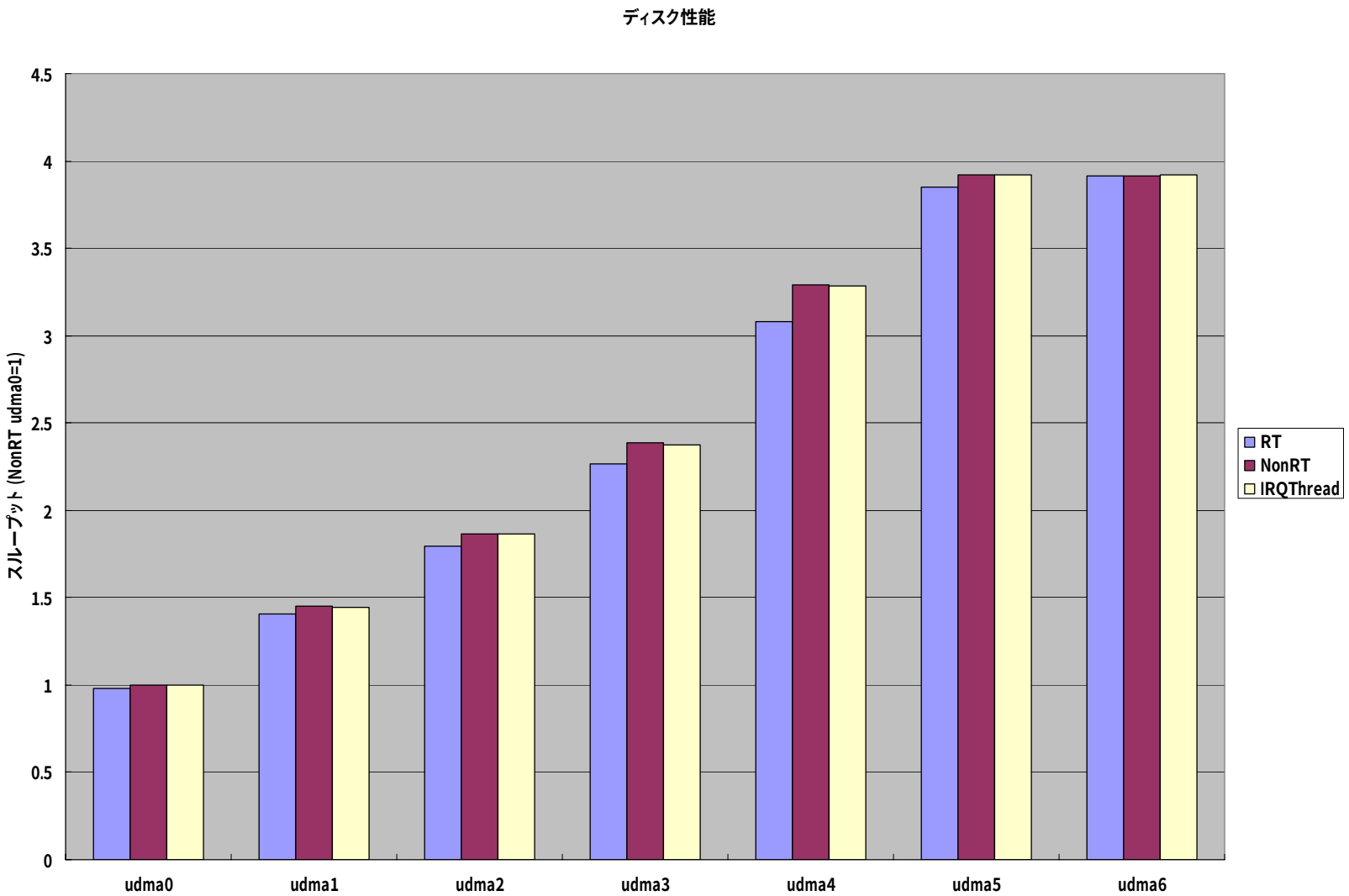




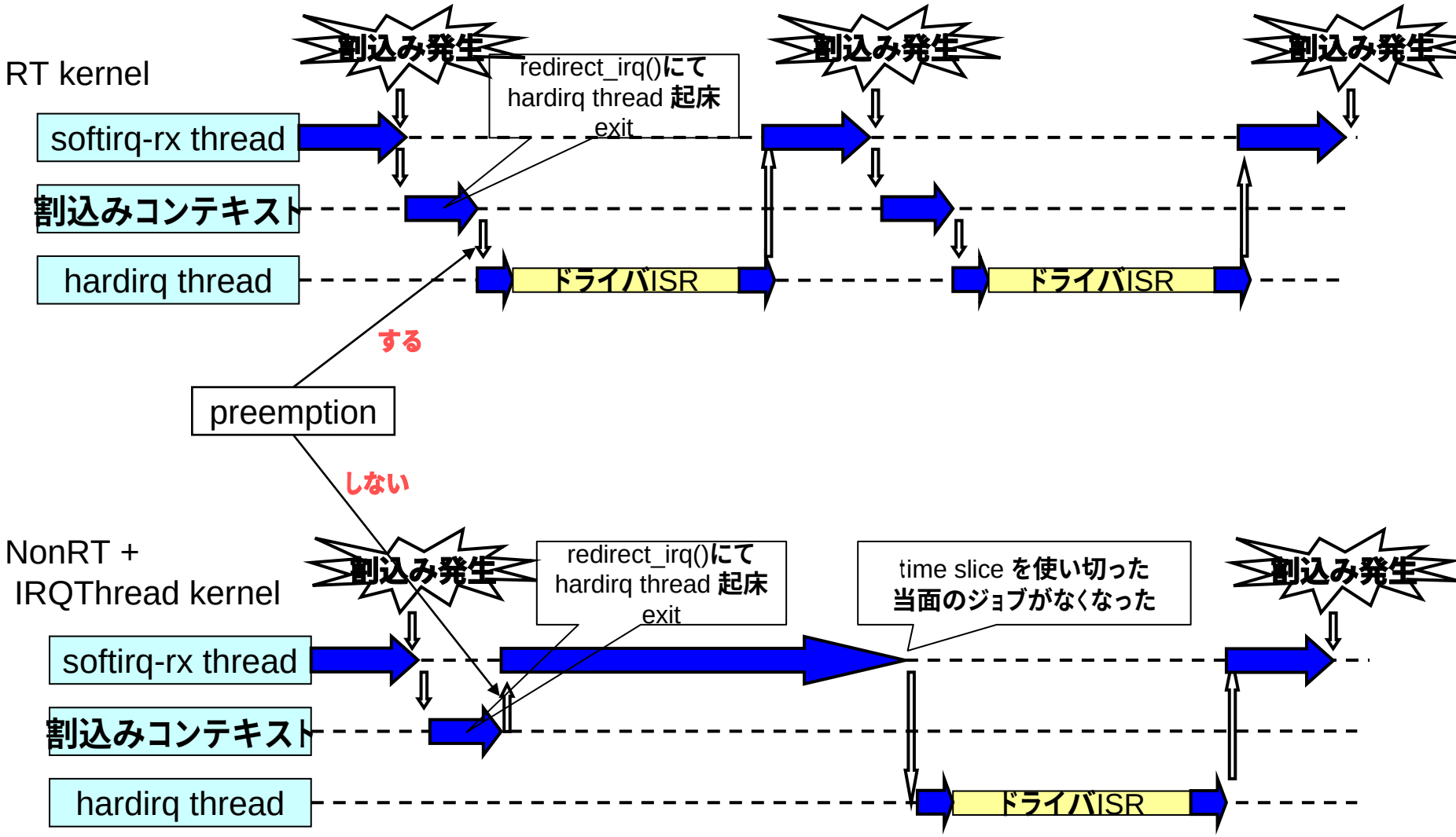
スループット ー ネットワーク性能



スループット ー ディスク性能



NonRT + IRQThead カーネルの性能が良いのは…



まとめ

- **RT vs NonRT**
  - lpar**割込み応答時間は負荷の種類に依存**
    - HDD → RT**版優位**
    - Network → **大差なし？ 若干**RT**版が悪い**
      - **元々遅延が少ない場合は**、IRQthread**へのcontext switch オーバヘッドが効く**
  - **スループット**
    - hdparm → RT**版健闘最悪でも ▼7%**
    - netperf → RT**版特に受信が ▼9%、送信は▼18%**
      - **何パーセントなら許容範囲？**

- **NonRT + IRQTheadカーネル**
  - netperf**送信時には有利に働いた可能性**
    - **余計**?**な割込みは後まわし**
    - **コントローラの**FIFO**に溜まったパケットを一括処理？**
  - **但し**、latency**には悪い影響を与えそう**

  - **疑問**：**そもそも**PREEMPT_NONE **と** PREEMPT_HARDIRQS**は両立するよう意図されているのだろうか？**

今後の課題

- **カーネルバージョンアップ**
  - 2.6.20 and/or 2.6.21
    - RT**パッチ**latest: patch-2.6.20-rt2 as of 2/6/2007

- **RT性検証･性能向上**
  - **更なる性能測定**
  - **不要な** raw_spinlock_t ?
  - **ドライバ**

- **Linux Community/CELFへ**
  - **移植パッチ提供**
  - PowerPC **その他のプラットフォームへ**
    - powermac…

TOSHIBA

Leading Innovation >>>